Regret-optimal policies in absorbing semi-Markov decision processes with multiple constraints(The Development of Information and Decision Processes)

AUTHOR(S):

Kadota, Yoshinobu; Kurano, Masami; Yasuda, Masami

CITATION:

Kadota, Yoshinobu ...[et al]. Regret-optimal policies in absorbing semi-Markov decision processes with multiple constraints(The Development of Information and Decision Processes). 数理解析研究所講究録 2006, 1504: 87-94

ISSUE DATE:

2006-07

URL:

http://hdl.handle.net/2433/58497

数理解析研究所講究録
1504巻 2006年 87-94

Regret-optimal policies in absorbing semi-Markov decision processes with multiple constraints

門田良信 (和歌山大学教育学部) Yoshinobu Kadota[a]
蔵野正美 (千葉大学教育学部) Masami Kurano[b]
安田正實 (千葉大学理学部) Masami Yasuda[c]

[a]Faculty of Education, Wakayama University, Wakayama 640-8510, Japan
[b]Faculty of Education, Chiba University, Chiba 263-8522, Japan
[c]Faculty of Science, Chiba University, Chiba 263-8522, Japan

Abstract

We consider a constrained regret-optimization problem for semi-Markov decision processes. The expected regret-utility of the total reward is minimized subject to multiple expected regret-utility constraints and the planning horizon is a reaching time to a given absorbing subset. By introducing a corresponding Lagrange function, a saddle point theorem is given. The existence of a constrained optimal policy is characterized by optimal action sets specified with a parametric utility*.

1. Introduction and notation

In decision making, it may be more appropriate to evaluate each decision or policy under a regret-optimality criterion. In our preceding work[12], we consider the general regret-utility problem for absorbing semi-Markov decision processes(semi-MDPs), in which the expected utility of the total reward earned until the stopping time is minimized. The regret-optimal policy is characterized by the corresponding optimality equation.

In this paper, we are concerned with the constrained optimization problem for the same model as [12]. In fact, it often occurs, in a social life or in a business that that we want to maximize the reward under several utility functions. For example, in the group decision making with different utility functions each player wants to maximize the reward under his own utility function. In such a case, not only one type of expected utility but other types are desired to keep higher than some given bound.

Here, we consider the constrained regret-optimization problem for semi-MDPs in which the expected regret-utility of the total reward earned until the reaching time to a given absorbing subset is minimized subject to multiple expected regret-utility constraints and the objective is to show that the Lagrange approach to the utility-constraints case is made successfully. In fact, by introducing a corresponding Lagrange function, a saddle point theorem is obtained and the existence of a constrained optimal policy is proved. Also a constrained optimal policy is characterized by optimal action sets specified with the parametric utility.

Keywords: Semi-Markov decision process; Utility constraint; Lagrange technique; Saddle point; Optimal policy.

In the same way with [12], we do not impose a special condition on the regret-utility functions, expecting to enlarge the practical application of the optimization problem. For the utility discussions for MDPs and constrained MDPs, refer to [5, 6, 8–11, 13] and their references. In remainder of this section, a constrained regret-utility optimization problem is formulated for the absorbing semi-MDPs model.

A semi-MDP is specified by the next five components:

(i) a countable state space: $S = \{0, 1, 2, \cdots\}$,

(ii) a finite action space: $A = \{1, 2, \cdots, m\}$,

(iii) transition probability distributions: $\{(p_{ij}(a); j \in S) \mid i \in S, a \in A\}$,

(iv) distribution functions $\{F_{ij}(\cdot|a) \mid i, j \in S,\ a \in A\}$ of the time between transitions,

(v) an immediate reward r and a reward rate d which are functions from $S \times A$ to $\boldsymbol{R}_+$, where $\boldsymbol{R}_+ = [0, \infty)$.

When the system is in state $i \in S$ and action $a \in A$ is taken, then it moves to a new state $j \in S$ with the sojourn time τ, and the reward $r(i, a) + d(i, a)\tau$ is obtained, where the new state j and the sojourn time τ are distributed with $p_{i\cdot}(a)$ and $F_{ij}(\cdot|a)$ respectively. This process is repeated from the new state $j \in S$.

The sample space is the product space $\Omega = (S \times A \times \boldsymbol{R}_+)^\infty$. Let X_n, Δ_n and τ_{n+1} be random quantities such that $X_n(\omega) = x_n$, $\Delta_n(\omega) = a_n$ and $\tau_{n+1}(\omega) = t_{n+1}$ for all $\omega = (x_0, a_0, t_1, x_1, a_1, t_2, \cdots) \in \Omega$ and $n = 0, 1, 2, \cdots$. Let $H_n = (x_0, a_0, t_1, \cdots, x_n)$ be a history until time n. A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of conditional probabilities $\pi_n = \pi_n(\cdot \mid H_n)$ such that $\pi_n(A \mid H_n) = 1$ for all histories $H_n \in (S \times A \times \boldsymbol{R}_+)^n \times S$. The set of all policies is denoted by Π. A policy $\pi = (\pi_0, \pi_1, \cdots)$ is called stationary if there exists a function $f : S \to A$ such that $\pi_n(\{f(X_n)\} \mid H_n) = 1$ for all $n \geq 0$ and $H_n \in (S \times A \times \boldsymbol{R}_+)^n \times S$. Such a policy is denoted by f^∞ or f.

For any $\pi \in \Pi$, we assume that

(i) $Prob(X_{n+1} = j|\ X_0, \Delta_0, \tau_1, \cdots, X_n = i, \Delta_n = a) = p_{ij}(a)$ and

(ii) $Prob(\tau_{n+1} \leq t|\ X_0, \Delta_0, \tau_1, \cdots, X_n = i, \Delta_n = a, X_{n+1} = j) = F_{ij}(t|a)$

for all $n \geq 0$, $i, j \in S$ and $a \in A$.

For any Borel set D, we denote by $P(D)$ the set of all probability measures on D. From (i) and (ii), we can define the probability measure $P_\pi^\nu \in P(\Omega)$ with an initial distribution $\nu \in P(S)$ with $Prob(X_0 = i) = \nu(i), i \in S$ and $\pi \in \Pi$ by a usual way.

For any subset $J_0 \subset S$, called as absorbing set, let

$$N := \min\{n > 0 | X_n \in J_0\}, \quad \text{where} \quad \min \emptyset = \infty. \tag{1.1}$$

The present value $\{\widetilde{D}_\ell : \ell = 1, 2, \cdots\}$ and the total lapsed time $\{\tilde{\tau}_\ell : \ell = 1, 2, \cdots\}$ of the process $\{X_n, \Delta_n, \tau_{n+1} : n = 0, 1, 2, \cdots\}$ until the ℓ-th time are defined respectively by

$$\begin{aligned} &\widetilde{D}_\ell := \textstyle\sum_{n=0}^{\ell-1} (r(X_n, \Delta_n) + \tau_{n+1} d(X_n, \Delta_n)) \quad \textit{and} \\ &\tilde{\tau}_\ell := \textstyle\sum_{n=1}^{\ell} \tau_n, \quad (\ell \geq 1). \end{aligned} \tag{1.2}$$

Let G, H_i $(i = 1, 2, \cdots, k) : \boldsymbol{R}_+ \times \boldsymbol{R}_+ \to \boldsymbol{R}$ be Borel-measurable functions, which will be called regret-utility functions as describing the general evaluation between the target value and the present value.

For any given threshold vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k) \in \boldsymbol{R}^k$ and constraint-target vector $h = (h_1, h_2, \cdots, h_k) \in \boldsymbol{R}^k$, let

$$V(\nu, \alpha, h) := \left\{ \pi \in \Pi \,\middle|\, E_\pi^\nu[H_i(h_i\tilde{\tau}_N, \widetilde{D_N})] \leq \alpha_i \text{ for all } i \ (1 \leq i \leq k) \right\}, \tag{1.3}$$

where $E_\pi^\nu[\cdot]$ is the expectation w.r.t. $P_\pi^\nu[\cdot]$. If $\nu(\{i\}) = 1$ for some $i \in S$ we write P_π^ν by P_π^i and E_π^ν by E_π^i.

Then, for a constant g^*, called a target value, our problem is given in the following.

Problem A: Minimize $E_\pi^\nu[G(g^*\tilde{\tau}_N, \widetilde{D_N})]$ subject to $\pi \in V(\nu, \alpha, h)$.

The optimal solution $\pi^* \in V(\nu, \alpha, h)$ of Problem A, if it exists, is called a ν-constrained regret-optimal policy or simply an optimal policy. For any $\nu \in P(S)$, let

$$\varphi(\nu) := \left\{ P_\pi^\nu \in P(\Omega) \,\middle|\, \pi \in \Pi \right\}.$$

Then, by a slight modification of the proof of Theorem 3.2 and 3.3 in V.S.Borkar[3], we have the following assertion.

Lemma 1.1 *For any $\nu \in P(S)$, $\varphi(\nu)$ is a convex and compact set in the weak topology.*

Regret-optimality is motivated by average optimality. Suppose a Markov chain corresponding to each stationary policy f is possitive recurrent and irreducible. Let any state $0 \in S$ be absorbing and $\tilde{\tau} = \min\{n \geq 1;\ X_n = 0\}$. Then we have from Theorem 7.5 of Ross[17], for a real number δ and a bounded function v,

$$\lim_{T\to\infty} \frac{1}{T} E_f(\sum_t v(X_t, \Delta_t) \,|\, 0) \leq \delta \quad \text{if and only if} \quad E_f[\delta\tilde{\tau} - \sum_t v(X_t, \Delta_t) \,|\, 0] \geq 0. \tag{1.4}$$

Letting $\delta = g^*$ the average optimal value in (1.4), maximization of the first term is equivalent to minimization of the second term which is the case $G(x, y) = x - y$. And (1.4) is also valid for $\delta = \alpha_j$. Thus, Problem A is closely related to the utility genralization of a constrained average optimal problem.

In Section 2, the saddle point statement for Problem A will be described, for the purpose of obtaining the existence of a ν-constrained regret-optimal policy. In Section 3, characterization of this optimal policy will be given.

2. Saddle point theorem for constrained semi-MDP

Now we discuss the saddle point theorem for Lagrangian associated with Problem A. For any initial distribution $\nu \in P(S)$, Lagrangeian L^ν is defined by

$$L^\nu(\pi, \lambda) := \sum_{i=1}^k \lambda_i(\alpha_i - E_\pi^\nu[H_i(h_i\tilde{\tau}_N, \widetilde{D_N})]) - E_\pi^\nu[G(g^*\tilde{\tau}_N, \widetilde{D_N})] \tag{2.1}$$

for any $\pi \in \Pi$ and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \in \boldsymbol{R}_+^k$. Without any confusion, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \in \boldsymbol{R}_+^k$ will be written simply by $\lambda \geq 0$.

The following statement on saddle points can be proved similarly to that of Theorem 2 in Luenberger [15] at § 8.5. The proof is omitted.

Theorem 2. 1 (cf. [15]) *Suppose that there exist* $\pi* \in \Pi$ *and* $\lambda^* \geq 0$ *such that* L^ν *with* $\nu \in P(S)$ *possesses a saddle point at* π^*, λ^*, *i.e.,*

$$L^\nu(\pi, \lambda^*) \leq L^\nu(\pi^*, \lambda^*) \leq L^\nu(\pi^*, \lambda) \tag{2.2}$$

for each $\pi \in \Pi$ *and* $\lambda \geq 0$. *Then,* π^* *solves Problem A and is a* ν*-constrained regret-optimal policy.*

This theorem motivates us to obtain a sufficient condition for the existence of a saddle point associated with Lagrangian L^ν. We need the following assumptions.

Assumption 2. 1 (i) *There exist* M_1 *and* M_2 *such that* $0 \leq r(i,a) \leq M_1 < \infty$, $0 \leq d(i,a) \leq M_2 < \infty$ *for all* $i \in S$, $a \in A$.

(ii) *There exist* $L > 0$, $B > 0$ *such that* $L \leq \int_0^\infty t\, F_{ij}(dt|\, a) \leq B$ *for all* $i, j \in S$ *and* $a \in A$.

Assumption 2. 2 $K := \sup_{\pi \in \Pi} E_\pi^\nu(N) < \infty$.

Assumption 2. 3 *Regret-utility functions* G, H_i $(i = 1, 2, \cdots, k)$ *are all lower semi-continuous.*

Now we give sufficient conditions for Assumption 2. 2 to hold. Define $e(n)$, $n = 1, 2, \cdots$ by $e(n) = \sup_{i \in S} e_i(n)$, where $e_i(n) = \sup_{\pi \in \Pi} P_\pi^i(N > n)$. Then, it holds that $e(n+1) \leq e(n)$ and $e(n+m) \leq e(n)e(m)$ for all $m, n = 1, 2, \cdots$.

Proposition 2. 1 (cf. [12]) *Each of the following conditions* (i) *and* (ii) *satisfies Assumption* 2. 2 .

(i) $\sum_{n=1}^\infty e(n) < \infty$.

(ii) *There exist* $0 < \eta_0 < 1$ *and* $n_0 > 1$ *such that* $e(n_0) < 1 - \eta_0$.

Let, for each $\nu \in P(S)$ and $\pi \in \Pi$, define a class $\Phi(\nu)$ by

$$F_\pi^\nu(x, y) := P_\pi^\nu(\widetilde{\tau_N} \leq x, \widetilde{D_N} \leq y) \quad \text{and} \tag{2.3}$$

$$\Phi(\nu) := \left\{ F_\pi^\nu(\cdot, \cdot) \,\middle|\, \pi \in \Pi \right\}. \tag{2.4}$$

Here, with some abuse of notation, we define

$$L^\nu(F, \lambda) := \int_0^\infty \int_0^\infty g_\lambda(x, y)\, F(dx, dy) \tag{2.5}$$

for any $F \in \Phi(\nu)$ and $\lambda \geq 0$, where

$$g_\lambda(x, y) := \sum_{j=1}^k \lambda_j \left(\alpha_j - H_j(h_j x, y)\right) - G(g^* x, y). \tag{2.6}$$

Then, Lagrangian L^ν defined in (2.1) is obviously rewritten by $L^\nu(\pi, \lambda) = L^\nu(F, \lambda)$ with $F = F_\pi^\nu$. Thus, we have the following corollary.

Corollary 2. 1 *Let $\pi^* \in \Pi$ and $\lambda^* \geq 0$. $L^\nu(\cdot,\cdot)$ with $\nu \in P(S)$ possesses a saddle point at π^*, λ^* if and only if the following relation holds*

$$L^\nu(F,\lambda^*) \leq L^\nu(F^\nu_{\pi^*},\lambda^*) \leq L^\nu(F^\nu_{\pi^*},\lambda) \tag{2.7}$$

for all $F \in \Phi(\nu)$ and $\lambda \geq 0$. Then, π^* solves Problem A *and is a ν-constrained regret-optimal policy.*

Lemma 2. 2 *For any $\nu \in P(S)$, it holds that*

(i) *$\Phi(\nu)$ is convex and compact it the weak topology;*

(ii) *$L^\nu(\cdot,\lambda)$ is concave and upper semi-continuous for each $\lambda \geq 0$;*

(iii) *$L^\nu(F,\cdot)$ is convex and continuous for each $F \in \Phi(\nu)$.*

Proof. By Assumption 2. 1 and 2. 2 , we observe that

$$0 \leq E^\nu_\pi[\widetilde{\tau_N}] \leq BK \quad \text{and} \quad 0 \leq E^\nu_\pi[\widetilde{D_N}] \leq (M_1 + M_2 B)K$$

for all $\pi \in \Pi$. Also, $(\widetilde{\tau_N},\ \widetilde{D_N})$: $\Omega \longrightarrow \boldsymbol{R}_+ \times \boldsymbol{R}_+$ is continuous, so that from Lemma1. 1 , (i) follows. By using the Assumption 2. 3 , (ii) holds. For (iii), it follows clearly. □

From Lemma 2. 2 , Fan's minimax theorem(cf. [4]) can be applied to obtain the following lemma.

Lemma 2. 3 *It holds, for any $\nu \in P(S)$,*

$$\inf_{\lambda\geq 0}\max_{F\in\Phi(\nu)} L^\nu(F,\lambda) = \max_{F\in\Phi(\nu)}\inf_{\lambda\geq 0} L^\nu(F,\lambda) \tag{2.8}$$

Henthforth, the common value in the both side of (2.8) will be denoted simply by L^*. In order to prove the existence of a saddle point with (2.7), we need the following condition.

Slater Condition There exists a $\overline{\pi} \in \Pi$ such that

$$E^\nu_{\overline{\pi}}[H_i(h_i\widetilde{\tau_N}, \widetilde{D_N})] < \alpha_i \tag{2.9}$$

for all i $(1 \leq i \leq k)$.

Since $L^\nu(F^\nu_{\overline{\pi}},\lambda) \longrightarrow \infty$ as $\|\lambda\| \to \infty$ under condition (2.9), the convex function $\max_{F\in\Phi(\nu)} L^\nu(F,\lambda)$ is bounded from below, so that by (2.8) there exists $\lambda^* \geq 0$ such that

$$\max_{F\in\Phi(\nu)} L^\nu(F,\lambda^*) \leq L^*. \tag{2.10}$$

On the other hand, by Lemma 2. 3 , there exists $F^* \in \Phi(\nu)$ with

$$L^\nu(F^*,\lambda) \geq L^* \tag{2.11}$$

for all $\lambda \geq 0$. Thus, applying Corollary 2. 1 , the following main theorem follows.

Theorem 2. 2 *Under Slater condition* (2.9), *Lagrangian* $L^\nu(\cdot,\cdot)$ *with* $\nu \in P(S)$ *has a saddle point, i.e., there exists* $\pi^* \in \Pi$ *and* $\lambda^* \geq 0$ *satisfying* (2.2).

Also, from Theorem 2. 1 and 2. 2 , the following corollary holds.

Corollary 2. 2 *Under Slater condition* (2.9), *there exists a* ν*-constrained regret-optimal policy for* $\nu \in P(S)$.

3. Characterization of optimal policy

Now we will derive some theoretical results, which are useful to seek a ν-constrained regret-optimal policy. Letting $\nu \in P(S)$ and for each $\lambda \geq 0$, a policy $\pi \in \Pi$ is said to be g_λ-optimal if

$$E^\nu_{\pi^*}[g_\lambda(\widetilde{\tau_N}, \widetilde{D_N})] \geq E^\nu_\pi[g_\lambda(\widetilde{\tau_N}, \widetilde{D_N})]$$

for all $\pi \in \Pi$, where g_λ is defined in (2.6).

The following Lemma can be easily proved (cf. [2]).

Lemma 3. 1 *Let* $\overline{\pi} \in \Pi$ *and* $\overline{\lambda} = (\overline{\lambda}_1, \overline{\lambda}_2, \cdots, \overline{\lambda}_k) \geq 0$. *Then, Lagrangian* $L^*(\cdot,\cdot)$ *given in* (2.1) *has a saddle point at* $\overline{\pi}, \overline{\lambda}$ *if and only if the following* (i) – (iii) *holds:*

(i) $\overline{\pi}$ *is* $g_{\overline{\lambda}}$*-optimal;*

(ii) $\overline{\pi} \in V(\nu, \alpha, h)$;

(iii) $\sum_{i=1}^k \overline{\lambda}_i \left(\alpha_i - E^\nu_{\overline{\pi}}[H_i(h_i\widetilde{\tau_N}, \widetilde{D_N})]\right) = 0$.

For any Borel set X, we denote by $\boldsymbol{B}(X)$ the set of all bounded Borel measurable functions on X. We define an operator $U_\lambda(d)(c_0, c, x \mid i, a)$ for $d = (d_i; i \in S)$ with $d_i \in B(R_+^{k+2})$ providing that $c_0 \in \boldsymbol{R}$, $c = (c_1, c_2, \cdots, c_k) \in \boldsymbol{R}^k$, $x \in \boldsymbol{R}$ and $i \in S, a \in A$, by

$$\begin{aligned} & U_\lambda(d)(c_0, c, x \mid i, a) \\ (3.1) \quad = & \sum_{j \in J} p_{ij}(a) \int_0^\infty d_j\left(c_0 + g^* t, c + ht, x + r(i,a) + d(i,a)t\right) F_{ij}(dt|a) \\ & + \sum_{j \in J_0} p_{ij}(a) \int_0^\infty g_\lambda\left(c_0 + g^* t, c + ht, x + r(i,a) + d(i,a)t\right) F_{ij}(dt|a) \end{aligned}$$

where $J = S - J_0$ and

$$g_\lambda(c_0, c, x) = \sum_{j=1}^k \lambda_j(\alpha_j - H_j(c_j, x)) - G(c_0, x),$$

with $c_0 \in \boldsymbol{R}$, $c + ht = (c_1 + h_1 t, c_2 + h_2 t, \cdots, c_k + h_k t) \in \boldsymbol{R}^k$, $x \in \boldsymbol{R}$.

Now we define an optimal value function starting from the initial state $i \in S$ with $(c_0, c, x) \in \boldsymbol{R}_+^{k+2}$ by

$$\begin{aligned} & g_i^\lambda(c_0, c, x) \\ (3.2) \quad := & \inf_{\pi \in \Pi} E^i_\pi \left[\sum_{j=1}^k \lambda_j \left(\alpha_j - H_j(c_j + h_j\widetilde{\tau_N}, x + \widetilde{D_N})\right) - G(c_0 + g^*\widetilde{\tau_N}, x + \widetilde{D_N})\right]. \end{aligned}$$

Then we have the following by the same method of Theorem 2.1 in [12].

Lemma 3. 2 *For $\lambda \geq 0$, the set of optimal value functions $g^\lambda = \{g_i^\lambda; i \in S\}$ is given as a unique solution of the optimality equation;*

$$g_i^\lambda(c_0, c, x) = \min_{a \in A} U_\lambda(g^\lambda)(c_0, c, x \mid i, a) \tag{3.3}$$

for all $i \in S$ and $(c_o, c, x) \in \boldsymbol{R}_+^{k+2}$.

In order to determine an optimal policy, we define the set of λ-optimal actions $A^\lambda(c_0, c, x \mid i)$ by

$$A^\lambda(c_0, c, x \mid i) := \arg\min_{a \in A} U_\lambda(g^\lambda)(c_0, c, x \mid i, a),$$

where $g^\lambda = (g_i^\lambda; i \in S)$ is a unique solution of (3.3). Then we have the following theorem.

Theorem 3. 1 *For any $\nu \in P(S)$, a policy $\pi^* \in V(\nu, \alpha, h)$ is a ν-constrained regret-optimal policy if and only if there exists $\lambda^* \geq 0$ such that*

(i) $P_{\pi^*}^\nu \left(\Delta_t \in A^{\lambda^*}(g^* \widetilde{\tau}_t, h\widetilde{\tau}_t, \widetilde{D}_t \mid X_t)\right) = 1$ *for all $t \geq 0$, where $h\widetilde{\tau}_t = (h_1\widetilde{\tau}_t, \cdots, h_k\widetilde{\tau}_t)$;*

(ii) $\sum_{i=1}^{k} \lambda_i^* \left(\alpha_i - E_{\pi^*}^\nu \left[H_i(h_i\widetilde{\tau_N}, \widetilde{D_N}\right]\right) = 0.$

Proof. Applying the results of Theorem 2.1 in [12], it can be shown that π^* is g_{λ^*}-optimal if and only if the above (i) holds. So this theorem follows from Lemma 3. 1 . □

References

[1] Altman, E.; *Constrained Markov Decision Processes*, Chapman & Hall/CRC, 1999.

[2] Avriel, M.; *Nonlinear Programming, Analysis and Methods*, Prentice Hall, Inc., 1976.

[3] Borkar, V. S.; *Topics in Controlled Markov Chains*, Longman Scientific Technical, 1991.

[4] Borwein, J.M. and Zhuang, D.; On Fan's minimax theorem, *Math. Programming*, **34**, 232–244, 1986.

[5] Chung, K. J. and Sobel, M. J.; Discounted MDP's: Distribution functions and exponential utility maximization, *SIAM J. Control Optim.* **25**, 49-62, 1987.

[6] Denardo, E.V. and Rothblum, U.G.; Optimal stopping, exponential utility and linear programming, *Math. Prog.* **16**, 228–244, 1979.

[7] Fishburn, P.C.; *Utility Theory for Decision Making*, John Wiley & Sons, New York, 1970.

[8] Hinderer, K. and Waldmann, K.H.; The critical discount factor for finite Markovian decision processes with an absorbing set. *Math. Mech. Oper. Res.* **57**, 1–19, 2003.

[9] Howard, R.S. and Matheson, J.E.; Risk-sensitive Markov decision processes, *Manag. Sci.*, **8**, 356–369, 1972.

[10] Kadota, Y., Kurano, M. and Yasuda, M.; Discounted Markov decision processes with general utility, In *Proceeding of APORS' 94*, 330–337, World Scientific, 1995.

[11] Kadota, Y., Kurano, M. and Yasuda, M.; On the general utility of discounted Markov decision processes, *Int. Trans. Opr Res.* **5**(1), 27–34, 1998.

[12] Kadota, Y., Kurano, M. and Yasuda, M.; Regret-optimality in semi-Markov decision processes with an absorbing set, *Proceeding, The Sixth International Conference on Optimization: Techniques and Applications(ICOTA6)* Paper No.44, 1–14, 2004.

[13] Kadota, Y., Kurano, M. and Yasuda, M.; Discounted Marokov decision processes with utility constraints, *Computers and Mathematics with Applications*, **51**, 279–284, 2006.

[14] Lippman, S.A.; Maximal average reward policies for semi-Markov decision processes with arbitrary state and action space. *Ann. Math. Statist.*, **42**, 1717–1726, 1971.

[15] Luenberger, D.; *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[16] Pratt, J.W.; Risk aversion in the small and in the large, *Econometrica*, **32**, 122–136, 1964.

[17] Ross, S.M.; Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.

[18] Sennot, L. I.; Constrained discounted Markov decision chains, *Probability in the Engineering and Information Sciences*, **5**, 463–475, 1991.